UCRL-ID-151391

LDRD Report FY 03: Structure and function of regulatory DNA: a next major challenge in genomics

L. Stubbs

February 18, 2003

U.S. Department of Energy
Lawrence Livermore National Laboratory

DISCLAIMER

This work was performed under the auspices of the U. S. Department of Energy by the University of California, Lawrence Livermore National Laboratory under Contract No. W-7405-Eng-48.

UCRL-ID-151391

LDRD Report FY 03: Structure and function of regulatory DNA: a next major challenge in genomics

PI: Lisa Stubbs, Genome Biology Division, BBRP

I. Introduction and project motivation

With the human genome sequence now available and high quality draft sequences of mouse, rat and many other creatures recently or soon to be released, the field of Genomics has entered an especially exciting phase. The raw materials for locating the ~30-40,000 human genes and understanding their basic structure are now online; next, the research community must begin to unravel the mechanisms through which those genes create the complexity of life. Laboratories around the world are already beginning to focus on cataloguing the times, sites and conditions under which each gene is active; others are racing to predict, and then experimentally analyze, the structures of proteins that human genes encode. These activities are extremely important, but they will not reveal the mechanisms through which the correct proteins are activated precisely in the specific cells and at the particular time that is required for normal developmental, health, and in response to the environment. Although we understand well the three-letter code through which genes dictate the production of proteins, the codes through which genes are turned on and off in precise, cell-specific patterns remain a mystery. Unraveling these codes are essential to understanding the functions of genes and the role of human genetic diversity in disease and environmental susceptibility. This problem also represents one of the most exciting challenges in modern biology, drawing in scientists from every discipline to develop the needed biological datasets, measurement technologies and algorithms.

The LDRD effort that is the subject of this report was focused on establishing the basic technical and scientific foundations of a well-rounded program in gene regulatory biology at LLNL. The motivation for building these foundations was based on several drivers. First, with the sea-change in genomics, we sought to develop a new, exciting and foreward-thinking research focus for the LLNL genomics team, which could leverage their long-standing investment and expertise in mapping, sequencing, and genetics of human chromosome 19 (ch19). The starting point of this effort was the LLNL and Joint Genome Institute (JGI) genomics teams' success in sequencing mouse DNA related to ch19. This effort was the first comparative sequencing project to be conducted on such a large scale; a manuscript describing this work was published in *Science* shortly after this LDRD began (Dehal et al., *Science* 293: 104-111, 2001). Like protein-coding genes themselves, DNA elements that control gene expression are protected from structural change over evolutionary time, since their function is dependent on sequence integrity. By contrast, DNA that serves no function ("junk DNA") can accumulate mutations that change sequence content without biological consequence. As a result, the "junk DNA" of human and mouse are not at all alike in sequence; however, genes and regulatory elements (REs) of human, mouse, and other animals –as far removed from human and fish and birds-- are very similar in sequence and structure.

This difference between conserved, functional DNA and non-conserved "junk" DNA provides the key to the importance of comparative sequence alignments. In fact, aligning the DNA sequence of two distantly related genomes (like human and mouse) and searching for sequence similarity is the most efficient means of locating regulatory elements (REs) in genomic DNA (Rubin and Tall, *Nature*, 407, 265-269, 2000). This is especially critical, since genes and REs together comprise no more than 5% of the 3 billion base pairs that make up the human genome. Finding that 5% of "gold" in this huge code is a tremendous problem that has engaged biologists and computational scientists alike for several years. Since comparative sequence alignments are key to finding both genes and REs in human sequence, the ch19-mouse sequence alignment dataset therefore represented a treasure chest of potential information regarding the locations, structures, and evolution of those elements most likely to be involved in human health, development., susceptibility to environment and pathogens, and many other important features.

It was important for LLNL's future in genomics for us to invest in taking BBRP's team beyond its traditional investment in DNA sequencing technology and into new areas of genomics research. The biology of gene regulation is a broad field with applications ranging from human susceptibility and molecular medicine to managing microbes in the environment. Several new calls for FY03 funding have been or will be announced soon by DOE's Office of Biological and Environmental Research (OBER) in which regulatory genomics will play a significant role. In addition, in July, 2002, the National Institutes of Health announced a plan for new funding in the area of "functional genome annotation", focused precisely on the kind of studies that this LDRD project were founded upon. Although we entered this field with some strategic advantages, LLNL was poorly equipped and understaffed for successful competition in this important new field. The timely investment in this LDRD program has positioned us to compete in this rapidly growing field, which couples experimental biology to bioinformatics in a way that is familiar and well-suited to the environment of this Laboratory.

The primary, effort of this program has been to establish a computational pipeline for defining locations of genes and associated REs in ch19 and related mouse DNA, and experimental methods to test the functions of the predicted elements in live cells. However, REs work by serving as docking sites for specific protein complexes, comprised of cooperative groups of "transcription factor" (TF) proteins. The presence or absence of specific components of these complexes governs the "off" or "on" status of their target genes. An understanding of circuitry that controls gene expression will therefore require knowledge of all components –both DNA and protein -- of regulatory elements in their active and inactive states. We therefore also initiated a study of genes encoding the most prevalent class of TF proteins in human and mouse DNA, called zinc-finger (ZNF) proteins, which comprise a significant fraction of the genes encoded on ch19.

II. Scope

During the 18 months of LDRD funding, we focused on the following challenges:

A. Developing and acquiring better computational tools for identifying and analyzing genes and REs in the ch19 comparative sequence data set

B. Developing protocols for testing candidate RE function in cultured cells and parallelizing the methods for high-throughput use, and application of these methods to analyze REs uncovered from comparative analysis of ch19 and related mouse genes. We selected a region containing 20 genes, including 5 genes for which REs had been previously characterized (to serve as positive controls) and 15 ch19 genes for which human and mouse REs had not been tested, as an area of focus for this study.

C. Characterization of genes encoding predicted ZNF TF proteins: Although by analogy to the few known examples, we can presume that ZNF genes encode regulatory proteins, nothing is known of the total repertoire of TF proteins, how those TF repertoires differ between species, like human and mouse, and what gene "targets" are regulated by each predicted protein. The definition of ZNF proteins –which include predicted products of nearly 500 genes in human and mouse, or almost 2% of the total 30,000 genes – is a huge endeavor with great biological significance. Our studies represented a pilot to examine the coding potential and activities of a selected set of the more than 250 ZNF genes that have been predicted by computational means to be located in ch19.

II. Approach and Results

A. Goal 1: Building a team and a foundation for sustainable competitive funding

Our long-term goal is to develop a bioinformatics team that can design new and improved algorithms for comparative sequence analysis and RE characterization. But the purpose of work under this LDRD was to lay foundation for such an effort by bringing in existing tools and testing their efficiency using experimental

methods. Another focus of this LDRD was recruitment of the right individuals to do this kind of coupled experimental-computational work. Although all of the personnel we attracted during the period of LDRD funding are not yet officially in place the LDRD pilot project was exceptionally helpful at putting LLNL on the map in this new field and attracting excellent Ph.D. scientists to the Laboratory. Two postdoctoral researchers were recruited during the period of LDRD funding and two others are expected to arrive at LLNL this Spring, 2003.

Another major goal was to develop the methods, infrastructure, and preliminary results to permit us to be competitive for long-term funding. This LDRD funding was instrumental in our successful competition for new funding from D.O.E./OBER, which was awarded after the LDRD work was completed in October, 2002. The newly funded project is a large, multipart collaboration with one of the collaborators we cultivated during the LDRD period, Dr. Barbara Wold at Caltech. The collaborative project totals more than $2 million per year in new funding for the two participating institutions and builds directly on results obtained during our LDRD.

B. Goal 2: Developing a pipeline for regulatory element (RE) discovery:

We aligned sequence from ch19 and related mouse regions to identify sequence elements with properties consistent with RE structure and function. This involved the application and adaptation of a suite of existing comparative sequence analysis algorithms to identify conserved sequences and correlate their positions with genes, transcription-factor binding sites, CpG-islands, and other features known to be associated with promoter and enhancer sites. To acquire these tools we developed collaborative ties to outside groups who have already invested significant effort in this computational problem. These collaborators included Dr. Michael Zhang, a computer scientist and group leader at the Cold Spring Harbor Laboratory, NY; Dr. Barbara Wold, at California Institute of Technology, Pasadena, CA; Drs. Inna Dubchak, Eddy Rubin, Lior Pachter and Ivan Ovcharenko at Lawrence Berkeley Laboratory/UC Berkeley and others in the bioinformatics community. We tested all of these computational tools on two different 1 million base pair (1Mb) regions of ch19, and compared the results of all methods. The predictions of each type of method were tested experimentally and results were compared. This process permitted us to evaluate the best methods for predicting specific types of elements so that we could weight the various input predictions accordingly in our pipeline. We also evaluated the different parameter settings that could be used with each algorithm to assess the accuracy of each in predicting known REs. These analyses provided the baseline data for a new pipeline that will be created and tailored to our project needs in future, separately funded studies.

C. Goal3: High-throughput testing of RE function *in vitro*

Comparative analysis of ch19 yielded more than 3000 distinct candidate regions that have characteristics suggesting they function as RE sequences. To analyze this large number of elements, we adapted existing, tried-and-true protocols to high-throughput application. We focused on streamlining methods including (a) PCR and cloning candidate REs into appropriate “reporter gene” plasmid vectors, (b) transfection into human and mouse cell lines (4-6 cell lines per species, selected for efficiency and suitability in expression measurement experiments), and (c) and measurement of reporter gene expression. For practical/tactical reasons, we focused specific on testing promoter activities. Promoters are sequences that contain information that direct RNA polymerase and transcription machinery to the start-site of a gene. All genes contain at least one promoter, and many genes have alternative start sites and associated promoters. Since ch19 contains about 1500 genes, there is at least that number of promoters in that chromosome. The reporter gene assay (RGA) --in which a chromogenic or fluorescent reporter protein (e.g. luciferase) is expressed if an active promoter or enhancer is cloned into the vector (**Fig. 1**) – is the

standard methods for regulatory biologists and vectors, cell lines and other materials are readily available

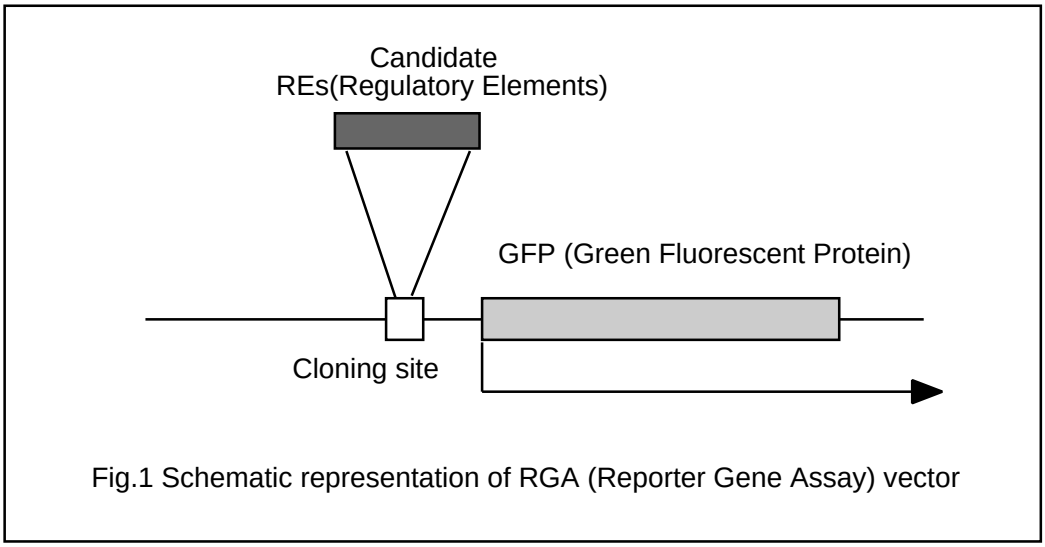

Fig.1 Schematic representation of RGA (Reporter Gene Assay) vector

from commercial sources. Our challenge was to turn the RGA methodology into a high throughput pipeline, multiplexing the steps involved in PCR, cloning, transfection, and reporter assays. REs that can function *in vivo* as a promoter element will guide the transcription machinery to start of the luciferase gene inserted in the adjacent position, a result ("glowing cells") that can be assayed easily and quantitatively in 96-well microtiter plates using a detection device that is available in our laboratory.

We used these methods to measure promoter and enhancer (expression stimulating) activity of REs associated with a set of 15 ch19 genes and 5 controls (genes with known promoters). The tests were conducted in different cell types and under several sets of experimental conditions in order to develop a robust protocol that is now ready for high-throughput use.

D. Goal 4: Classifying tested REs by linking each to specific gene expression patterns

In support of these studies we catalogued sites of expression for the genes located next to the promoters selected for testing. These studies link REs with specific structure and sequence to gene activity in specific cell types in living tissues. We measured mouse gene expression by *in situ* hybridization (ISH) in thin sections of whole newborn animals on microscope slides, to permit expression to be monitored on the level of the individual cell. Because of the newborn mouse's small size, we can complete a "whole-body gene

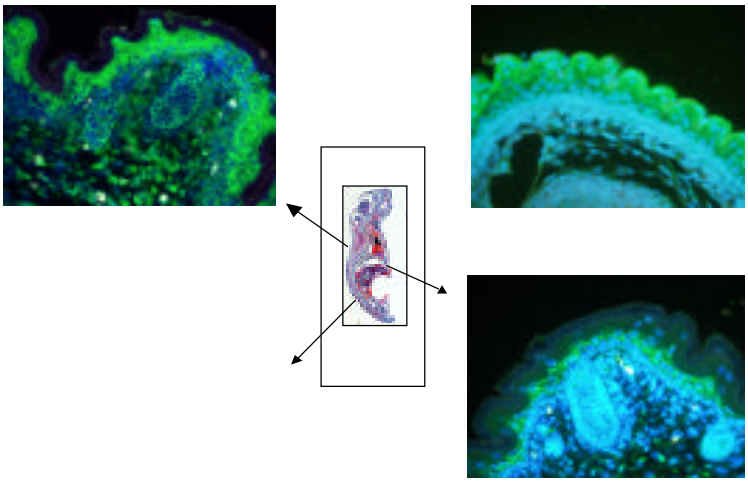

Figure 2. Adjacent 5 micron-thick saggital sections taken from a 3 day-old mouse (center) were hybridized with fluorescently-labeled gene-specific probes to detect activities of four genes from ch19. Gene expression shows up in these experiments as glowing green against a blue or red-stained (bottom left panel) background stain that is applied to show structure of the tissues. Each panel represents results obtained with a different ch19 gene. All four genes are active in the skin, but ISH shows that the genes are expressed in different cell layers and cell types.

expression scan" and catalog all expressing cells simultaneously. For example, four genes from a 1Mb test region of ch19 are expressed in the skin. But ISH experiments show us that the four genes are active in different layers and cell types in the skin **(Fig.2)**. We also established "tissue array" technology to study gene expression in adult tissues of humans and mice. To make tissue arrays, thin but deep cores of fixed tissue are placed into paraffin blocks; 5-10 micron-thick slices are then taken through the core. The result is that small dots of thin-sectioned tissue are arrayed on a slide, ready to be hybridized with fluorescent probe to detect locations of specific RNA. These studies permit us to classify the RE sequences according to their gene-controlling function in specific cells in living tissue

E. Cataloguing and characterizing species-specific transcription factor protein genes

One of the most intriguing findings arising from comparison of ch19 and mouse DNA was that certain types of genes differ greatly between the two species. Most human and mouse genes are very similar, and related to each other in a 1:1 fashion; that is for almost all of the 30,000 human genes one can define a clear mouse counterpart with similar structure and function. However, certain classes of genes have changed significantly in number and type over mammalian evolution, so that 1:1 relationships cannot be drawn between species. These genes have changed primarily through the generation of duplicate gene copies, which, over time, have changed in sequence to acquire new functions. Virtually all of the gene repertoire differences between human and mouse are due to this type of gene duplication mechanism.

Our studies in ch19 showed that one class of genes that has changed most rapidly in recent evolution are those that encode a particular type of transcription factor protein, the zinc-finger genes. Since the function of ZNF proteins is to orchestrate the activities of other genes, that fact that humans, mice and other mammals have different repertoires of such proteins indicate that gene regulatory networks have changed in species-specific ways. This finding may be a key to understanding the differences between species, and certainly has great potential to help us understand the way gene regulation works.

Although DNA sequence analysis shows clearly that mammalian genomes contain hundreds of ZNF genes, almost nothing is known about them or their protein products. Which genes are present in human but not mice, and vice versa? How are the existing human and mouse genes related to each other? Do all the genes encode functional proteins, and if so, in which cell types are they made? Which "target

genes" are orchestrated by the action of each ZNF protein? In this LDRD pilot study we characterized related ZNF gene sets in human and mice, clarified evolutionary relationships between the duplicated genes, and began work to clarify the functions of the species-specific genes. We discovered that despite their close evolutionary ties and their genomic and biological similarity, different ZNF genes exist even between mice and rats. The change in ZNF gene repertoire therefore appears to track speciation in a potentially interesting way. We predicted and confirmed structures of the genes and assessed their sites of activities in human and mouse tissues. We are carrying on this work with the motive of discovering which genes the ZNF proteins regulate and how gain or absence of specific genes affects biological pathways. The results of this pilot study have been summarized in a manuscript that is presently under review at the journal, *Genome Research.*

F. Summary

This LDRD project, which was funded for 18 months, has permitted us to build the infrastructure, methodologies, and research team required to launch a new program in regulatory genomics. We have used a set of genes from two human chromosome 19 regions to test different methods and approaches and to establish robust protocols that can be applied accurately in a high-throughput format. The preliminary results obtained from this pilot project were instrumental in boosting our competitiveness for funding in a larger, ambitious 5-year program that is now sponsored by DOE/OBER. We expect additional funding from the NIH to grow out of this LDRD pilot within the next year.